# ENFR: EVIDENCE NETWORKS WITH FLOW-MATCHED REFERENCES FOR AMORTIZED SIMULATION-BASED MODEL INFERENCE

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

The evidence, $\mathcal{Z} = p(x|\mathcal{M})$, of a data-generating model $\mathcal{M}$ is a powerful tool in Bayesian frameworks: it enables model inference, model misspecification tests, and goodness-of-fit assessment as an alternative to standard discrepancy statistics, e.g. $\chi^2$. While versatile, the evidence is notoriously difficult to compute, even when the likelihood is available. In both likelihood-based and simulation-based inference, estimating $\mathcal{Z}$ provides an immediate solution to high-dimensional goodness-of-fit or model misspecification tests: using $\mathcal{Z}(x)$ as the test statistic induces a minimum-volume level-$\alpha$ test. Such tests are essential to assess model fidelity, which is necessary to trust simulation-based parameter inference. Existing learned approaches largely target Bayes factors, which only measure evidence ratios. Instead, we provide accurate, fast, and amortized estimates of the absolute evidence. We present ENFR: *Evidence Networks with Flow-matched References*. Building on the Evidence Network framework of Jeffrey & Wandelt (2024), the proposed method first learns Bayes factors with respect to a flow-matched reference, from which the absolute evidence is recovered. We show improvement in estimation of $\mathcal{Z}$ over the flow-based baseline in all applications considered and the best performance of all density estimation methods considered in our benchmarks. We furthermore introduce and provide the proof for an architecture that jointly learns the individual evidences of a set of models $\{\mathcal{M}_i\}$. We demonstrate the performance and scaling behavior of these approaches on data from a variety of scientific applications.

## 1 INTRODUCTION

Science is fundamentally tasked with modeling nature from first principles, and increasingly this modeling is done with high-fidelity simulators that produce samples from a likelihood that cannot itself be evaluated (e.g. Agostinelli et al., 2003; Mesinger et al., 2011; Villaescusa-Navarro et al., 2021). Simulation-based inference (SBI) has changed how such models are confronted with data: instead of requiring a tractable likelihood, it enables likelihood-free inference directly from simulator samples (Papamakarios & Murray, 2016; Alsing et al., 2018; Brehmer et al., 2020; Cranmer et al., 2020). These methods have already unlocked information in the data that was previously inaccessible (e.g. von Wietersheim-Kramsta et al., 2025; Thomsen et al., 2026, for recent scientific applications), and are being readily adopted across scientific disciplines.

The focus of SBI so far has been parameter estimation. What we ultimately care about in science, however, is assessing the relative fits of models when confronted to data. In a Bayesian framework this is measured by the evidence $p(x|\mathcal{M})$, which is the probability that the observed data $x$ has been produced by a physical or statistical model $\mathcal{M}$. In parameterized settings it is common to express the evidence as the marginal likelihood using the model parameters $\theta$:

$$p(x|\mathcal{M}) = \int p(x|\theta, \mathcal{M}) p(\theta|\mathcal{M}) d\theta, \tag{1}$$

where $p(\theta|\mathcal{M})$ denotes the prior on the parameters and $p(x|\theta, \mathcal{M})$ is the model likelihood. The evidence serves this purpose in two ways. For model comparison, we are interested in the Bayes factor $\log K = \log \frac{p(x|\mathcal{M})}{p(x|\mathcal{M}_0)}$ with respect to some reference model $\mathcal{M}_0$. Beyond model comparison,



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the absolute evidence enables powerful tests of the data such as model misspecification tests, and an alternative to $\chi^2$ goodness-of-fit tests in high dimensions that does not require summary statistics.

Estimating the evidence has not been the focus of SBI, and the methods to do so are correspondingly less developed. Traditionally, computing the evidence has been a daunting task as the naive estimation of the integral is quickly intractable in high dimensions. Computationally heavy methods such as nested sampling (e.g. Feroz et al., 2009; Handley et al., 2015; Speagle, 2020) have been developed for these cases, but they require the ability to evaluate the likelihood, and even then have seen limited use in scientific applications due to computational infeasibility. How to evaluate the model evidence and Bayes factors when the likelihood is intractable but a simulator is available has remained a largely underexplored topic.

Jeffrey & Wandelt (2024) presented the application of neural network based density ratio estimators to learn Bayes factors in this regime. We build on this previous work and introduce ENFR, Evidence Networks with Flow-matched References, which combines flow-based models and evidence networks: using a generative flow as the reference model, we obtain not only Bayes factors but precise estimates of the absolute evidence. We furthermore introduce a new loss function to jointly estimate the Bayesian evidence and Bayes factors of multiple models in Section 3.

Alongside the method, we introduce an evaluation framework built on scientifically relevant applications from a range of disciplines in Section 4. It includes cases with a known likelihood, where the evidence can be estimated with nested sampling and our estimates can therefore be tested accurately against a ground truth, as well as simulation-based cases where only the calibration of the Bayes factors can be assessed. Across these applications, ENFR performs best of all learned estimators considered.

## 2 Related work

**Generative models for evidence estimation** Flow based generative models can both obtain samples from the target distribution and evaluate its density. In recent years, generative models such as flow matching (Lipman et al., 2023) or score-based diffusion (Song & Ermon, 2019; Song et al., 2021) have been successfully pushed into very high-dimensional regimes. However, their primary success criterion is often sample quality. An exact evaluation of the density is often not an explicit goal and its quality is not a main concern. In cases where the estimated density is evaluated as a metric, it is usually maximized, without a "true" value to aim for or any interpretable measure for its quality. For evidence estimation, however, the evaluation of the density is the final goal without any concerns of sample quality. Given the optimization objective in those two generative models, the velocity field and the score, it seems clear that these models do not prioritize the evaluation of the likelihood.

**Flow matching** In this work we consider flow matching (Lipman et al., 2023), learning the velocity field $v_\phi = \frac{dx}{dt}$ when mapping from samples of a known noise distribution $z \sim p(z)$ to data samples $x$ with learnable parameters $\phi$ and $t$ parameterizing the path. Throughout this work we consider linear interpolants for the flow, i.e. $x_t = tx + (1-t)z$. The density can then be obtained via integration from the continuity equation (see Appendix B.1).

**Evidence Networks** Given a reference model $\mathcal{M}_0$, one can use classifiers to learn a density ratio, or equivalently, the Bayes factor $\log K$ (Jeffrey & Wandelt, 2024). This can be achieved through a variety of loss functions (Masnadi-shirazi & Vasconcelos, 2008). Jeffrey & Wandelt (2024) recommend an exponential loss function which is given by:

$$\mathcal{L}(x, m) = e^{(\frac{1}{2}-m)\mathcal{J}_\alpha(f_\psi(x))} \tag{2}$$

where $m \in \{0, 1\}$ denotes the model label (e.g. $m = 1$ if $x \sim p(x|\mathcal{M})$ and $m = 0$ if $x \sim p(x|\mathcal{M}_0)$) and $f_\psi$ is a neural network with trainable parameters $\psi$. $\mathcal{J}_\alpha$ denotes the l-POP transform defined as $\mathcal{J}_\alpha(x) = x + \mathrm{sgn}(x)|x|^\alpha$ which provides for more accurate computation of extreme evidence ratios than cross-entropy loss. We explicitly show the minimizer of this loss to be the Bayes factor in Appendix E. The hyperparameter $\alpha$ can be chosen freely but is recommended by Jeffrey & Wandelt (2024) to be set to $\alpha = 2$. We follow this recommendation and will refer to this architecture as an evidence network (EN). As we are fundamentally interested in evidences as opposed to

Bayes factors, we need to choose a reference model with known density to reconstruct an evidence: $\log q_\psi(x|\mathcal{M}) = \log K_\psi + \log p(x|\mathcal{M}_0)$.
The choice of the reference model can be critical for good evidence estimation: evidence networks suffer from a well-known failure mode in the case where the distance between the two distributions $p(x|\mathcal{M})$ and $p(x|\mathcal{M}_0)$ is very large (Rhodes et al., 2020). This is known in the literature as the density chasm problem. In the following we describe approaches to mitigate this issue.

**Telescoping ratio estimation** As a solution to the density chasm problem, an iterative procedure has been proposed by Rhodes et al. (2020), in which intermediate waymark distributions are constructed, to produce an array of reference distributions and to learn the Bayes factors (or equivalently, density ratios) between consecutive waymark distributions. Choosing a known noise density $p(z)$, enables the reconstruction of the initial density. While Rhodes et al. (2020) have applied this also to energy based modelling with a flow-based noise distribution they do not consider controlled density-estimation benchmarks in which the reconstructed target density can be compared directly to its known normalized pointwise density. Furthermore our work here differs in architecture.

**Time Score Matching (CTSM)** A related continuous formulation is given by Choi et al. (2022); Yu et al. (2025); Chen et al. (2025), in which neural networks learn the time score $s(x,t) = \partial_t \log p(x_t)$. These models are in training similar to flow-matching, i.e. the target and noise distribution are connected with some interpolant. The learned time score can then be integrated to obtain a density estimate: $\log q(x) = \log p(z) + \int_0^1 s(x,t)\,\mathrm{d}t$. We train these models using the conditional time score matching objective (CTSM) (for more details see Yu et al., 2025) with either the variance preserving interpolant (VP) or the Schrödinger bridge interpolant (SB) (we show the interpolants in Appendix B.3). We note that while these methods have been primarily formulated for density ratio estimation, we evaluate them here as density estimators since we are fundamentally interested in evidences rather than Bayes factors. In our experiments, we have found the variance of the target to be very high close to the data distribution and to be limiting for the performance of this method. We also note, that under imperfect training, the resulting probability density function is not guaranteed to be normalized as opposed to flow-based generative models. In our numerical experiments we compare with CTSM-VP, and CTSM-SB with a learned flow matching reference distribution, which we call CTSM-SB + FlowRef.

**Featurized Density Ratio Estimation** Another related approach is featurized density ratio estimation (Choi et al., 2021) which uses an invertible flow to learn a common feature space of samples from two densities and to approximate the density ratio in this feature space. Again our work differs as we are not only interested in density ratios, but also actual densities.

**Flow contrastive estimation** Gao et al. (2019) propose flow contrastive estimation, which jointly trains an energy-based model and a normalizing flow through a noise-contrastive binary-classification objective, with the flow serving as an adaptive noise distribution. The flow is brought iteratively closer to the data distribution. While they do present density estimation results, they do not fully benchmark their density estimation for pointwise ground-truth examples beyond two dimensions. They furthermore show that maximizing the likelihood and improving sample quality do not necessarily go together. Their method differs from ours both in architectural setup and training method.

## 3 ENFR: Evidence Networks with Flow-Matched References

**Evidence Networks with Flow-matched References: ENFR** Given that ENs are a powerful method for the estimation of the evidence if a fitting reference model is present, we propose to learn such a reference using a generative model (see Figure 1). We choose flow matching for this task. The EN with learnable parameters $\psi$ then only has to learn a correction $\log K_\psi$ from the flow estimate to the true underlying density by discriminating simulations from flow samples:

$$\log p(x|\mathcal{M}) = \log q_\phi(x|\mathcal{M}) + \log \frac{p(x|\mathcal{M})}{q_\phi(x|\mathcal{M})} \approx \log q_\phi(x|\mathcal{M}) + \log K_\psi. \tag{3}$$

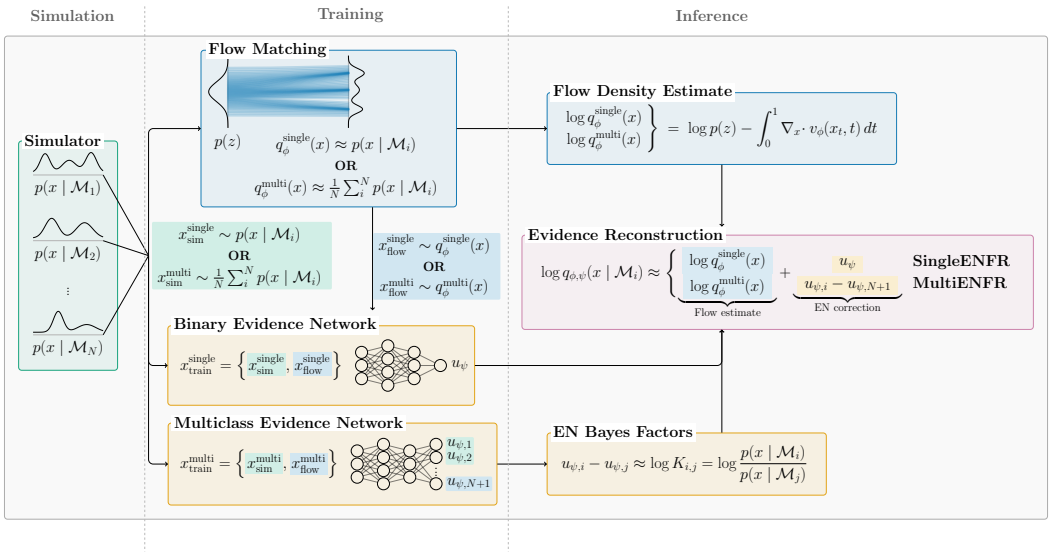



Figure 1: A visualization of our proposed SingleENFR and MultiENFR architectures. Given simulated data from one or multiple models, we train flow matching once over all models. In a second step, we use the simulated data and samples from the flow to train an evidence network classifier, which learns all Bayes factors. Those can be combined with the evaluated flow density to obtain the evidence under a given model.

In this setup, we can benefit from the fact that we only need accurate sampling from the flow to generate half of the training data for the evidence network. The flow density has to be evaluated only at inference time. We call this setup SingleENFR.

We show that when the flow is already a very good fit to the data and no informative density-ratio correction is available, the population-optimal constant correction is zero under balanced training in Appendix F. While we could use any evidence network loss function for the estimation of $\log K$, we follow the recommendation of Jeffrey & Wandelt (2024) and choose l-POP. We note that in a similar fashion, one can use a pre-trained flow as the base distribution for CTSM using the SB bridge which has been used in Chen et al. (2025). We also compare to that case in our experiments, calling it CTSM-SB + FlowRef.

**ENFR with multiple models (MultiENFR)** We further introduce a more streamlined way to obtain an array of Bayesian evidences for a set of multiple models based on the l-POP loss function which we call MultiENFR (see Figure 1). Given $N$ such models, we introduce the pairwise loss:

$$\mathcal{L}(x, m) = \frac{1}{N-1} \sum_{i \neq m} e^{-\frac{1}{2}(\mathcal{J}_\alpha(f_{\psi,m}(x)) - \mathcal{J}_\alpha(f_{\psi,i}(x)))}. \tag{4}$$

**Theorem 1** (Asymptotic recovery of Bayes factors)**.** *Consider $N \geq 2$ models $\{\mathcal{M}_i\}_{i=1}^N$ with evidences $p_i(x) \equiv p(x \mid \mathcal{M}_i)$, sampled with equal probability $p(\mathcal{M}_i) = 1/N$. Let $f_{\psi,i}(x)$ denote the $i$-th output of the multiclass evidence network and let $\mathcal{J}_\alpha$ be a fixed scalar link function, with transformed outputs $u_{\psi,i}(x) \equiv \mathcal{J}_\alpha(f_{\psi,i}(x))$.*

*Suppose the network is trained with the pairwise exponential loss $\mathcal{L}(x, m) = \frac{1}{N-1} \sum_{j \neq m} \exp[-\frac{1}{2}(u_{\psi,m}(x) - u_{\psi,j}(x))]$ (Equation 4). Assume that the densities $p_i(x)$ are positive almost everywhere on the common support, that the population risk is finite, that the transformed network class is sufficiently expressive to represent $u_i(x) = \log p_i(x) + c(x)$ for some model-independent function $c(x)$, and that asymptotic training reaches a global minimizer of the population risk.*

*Then every global population-risk minimizer satisfies $u_i^*(x) = \log p(x \mid \mathcal{M}_i) + c(x)$ almost everywhere, where $c(x)$ is independent of the model index. Consequently, for any pair of models $\mathcal{M}_i$ and*

$\mathcal{M}_j$,

$$\mathcal{J}_\alpha(f_{\psi^*,i}(x)) - \mathcal{J}_\alpha(f_{\psi^*,j}(x)) = \log \frac{p(x \mid \mathcal{M}_i)}{p(x \mid \mathcal{M}_j)} = \log K_{ij}(x). \tag{5}$$

*Proof.* See Appendix G. □

In order to handle the density chasm problem, we introduce one more class, which is given by synthetic data samples from a flow $q_\phi^{\text{multi}}$ with trainable parameters $\phi$. The generating flow is pretrained on simulated data from all the other models: $x \sim \frac{1}{N}\sum_{n=1}^{N} p(x|\mathcal{M}_n)$. We note that this is a fundamentally different flow density target than in SingleENFR, as the result is no longer an evidence estimate in itself. It is rather an estimate of the mixture of the evidences of many models. We use this flow reference as an additional model in our multiclass evidence network: there are therefore $N + 1$ classes to discriminate. We generate an equal share of training data from this pretrained flow. The learned Bayes factor with respect to the flow class along with the evaluated flow density then delivers the evidence $\log p(x|\mathcal{M}_i) = \log K_{i,N+1} + \log q_\phi(x)$.

# 4 RESULTS

In this section we benchmark the performance of our proposed ENFR architecture and compare them to other learned evidence estimates on a variety of different scientific applications.

## 4.1 ANALYTICAL EVIDENCE: GAUSSIAN MIXTURE MODEL

We begin with a Gaussian mixture model (GMM) where the true evidence is analytically known and we compare root mean square errors (RMSE) of our simulation-based estimates with respect to the ground truth. In the known ground truth case we generate data from a $d$-dimensional correlated Gaussian mixture model (GMM) with three components. The details of the experimental setup are given in Appendix A.1.

Figure 2 shows that in our scaling tests the SingleENFR setup consistently outperforms the other tested models. We also find that CTSM is highly outperformed by all other methods. Using CTSM with a flow matched reference model, does not improve the flow significantly in all combinations tested here. A comparison of the RMSE values of the evidence estimates with respect to the ground truth is given in Table 11. We also find that our model learns only minor improvements starting at 512 dimensions.

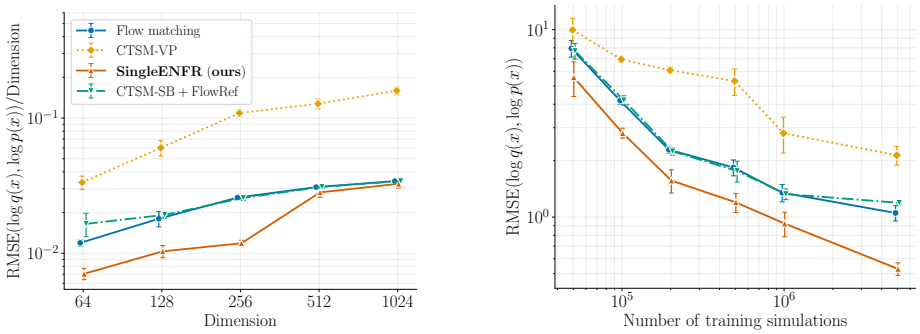
## 4.2 ANALYTICAL LIKELIHOOD APPLICATIONS

In this section we present examples where the true underlying likelihoods are known, such that the evidences can be estimated with established likelihood-based frameworks. We benchmark the models by comparing the simulation-based evidences to the likelihood-based estimates via the root mean square error (RMSE) of the log probabilities. For our likelihood-based estimates we run nested sampling using the dynesty library Speagle (2020).

### 4.2.1 BARYONIC ACOUSTIC OSCILLATIONS

We further test on a scientific example from cosmology. In the early Universe, sound waves propagated through the tightly coupled photon baryon plasma, driven by the competition between gravity and radiation pressure. When photons decoupled from baryons, these waves froze in and imprinted a characteristic length scale, the sound horizon $r_d$, on the correlations between cosmic structures. Baryonic acoustic oscillations (BAO) thus provide a *standard ruler*: measuring its apparent size along and across the line of sight at different redshifts constrains the expansion history $H(z)$, and therefore the nature of the dark energy that drives the late-time accelerated expansion.

In the standard cosmological model, ΛCDM, dark energy is a cosmological constant with a fixed equation of state $w = -1$. Whether this description is sufficient is currently an open question, and




(a) (b)

Figure 2: Scaling of density-estimation accuracy with (a) data dimensionality and (b) number of training simulations on the Gaussian mixture model of Appendix A.1, for which the true density is known analytically. We report the RMSE between estimated and true log-density over 200 held-out test points. In (a) the simulation budget is fixed at $5 \times 10^6$; in (b) the dimensionality is fixed at 64. The line shows the mean across three trained models and the errorbars show the estimated standard deviation. SingleENFR (orange) improves on the flow it is built on (blue) at every point and is the most accurate of all estimators considered, while CTSM-VP (yellow) is consistently the least accurate.

Table 1: RMSE of our learned log-evidence estimates against the likelihood-based log-evidence estimate over 100 held out BAO test simulations (lower is better). Our Models SingleENFR and MultiENFR consistently outperform both the flow and CTSM-SB with a flow reference.

| Simulation | flow | SingleENFR (ours) | MultiENFR (ours) | CTSM-SB + FlowRef |
|---|---|---|---|---|
| $\Lambda$CDM | 1.34 | 0.53 | **0.52** | 2.74 |
| $w$CDM | 1.40 | 0.76 | **0.67** | 2.46 |
| $w_0w_a$CDM | 1.16 | 0.83 | **0.81** | 3.79 |
| $\nu$CDM | 1.41 | 0.47 | **0.46** | 1.77 |
| $\Omega_k$CDM | 1.52 | **0.54** | 0.78 | 2.29 |

different extensions have been proposed: a constant but free equation of state ($w$CDM), a time-evolving one parameterized as $w(a) = w_0 + w_a(1-a)$ ($w_0w_a$CDM), as well as one-parameter extensions that free the neutrino mass sum ($\nu$CDM) or the spatial curvature ($\Omega_k$CDM). Recent BAO measurements from DESI Abdul Karim et al. (2025) have renewed interest in dynamical dark energy, making a principled comparison of these models a timely application. We compare the evidences of these five models for BAO data vectors (more details in Appendix A.2).

Table 1 shows the resulting evidences and Bayes factors on 100 held out test simulations for the different cosmological models. We find that our models SingleENFR and MultiENFR are consistently outperforming both the flow, and CTSM-SB with the flow reference. Table 2 shows the evidence estimates of our best model, MultiENFR, on the state-of-the-art measurement DESI DR2 (Abdul Karim et al., 2025). The estimates on the data agree on the order of the nested sampling error bars.

#### 4.2.2 Time series from an ODE

In this example we apply our method to a 64 dimensional time series from an ordinary differential equation (ODE). The data consists of noisy measurements of the value $x(t)$ at the 64 evenly spaced time steps. We consider the harmonic ($\beta = 0, \gamma = 0$), damped($\beta = 0, \gamma \neq 0$), and Duffing ($\beta \neq 0, \gamma \neq 0$) oscillators:

$$\frac{\mathrm{d}^2x}{\mathrm{d}t^2} = -2\gamma\frac{\mathrm{d}x}{\mathrm{d}t} - \omega^2 x - \beta x^3. \tag{6}$$

Table 2: Evidence estimates using MultiENFR on the DESI DR2 observed data. For MultiENFR we also quote the estimate of the standard deviation over the ensemble of 5 members. We can reproduce the likelihood-based evidence estimates of the DESI DR2 data with our simulation-based framework within the ensemble scatter.

| Model | Likelihood-based analysis | MultiENFR | $\Delta$ | MultiENFR $\log K$ |
|---|---|---|---|---|
| $\Lambda$CDM | -13.99±0.16 | -14.07±0.34 | -0.08 | 0.00 |
| $w$CDM | -16.41±0.18 | -16.00±0.35 | 0.40 | -1.94 |
| $w_0w_a$CDM | -15.63±0.18 | -15.46±0.40 | 0.17 | -1.40 |
| $\nu$CDM | -14.06±0.16 | -14.01±0.30 | 0.04 | 0.06 |
| $\Omega_k$CDM | -15.88±0.17 | -15.72±0.23 | 0.16 | -1.65 |

At each time step we add noise $\epsilon \sim \mathcal{N}(0, (0.1)^2)$. The priors are given in Table 4.

Figure 3a shows that our MultiENFR model outperforms all other models tested here. Both MultiENFR and SingleENFR evidence estimates are significantly more precise than those of the flow. We present our full results in Table 5, where we find again that the combination of CTSM-SB with a flow reference has the highest RMSE, i.e. the worst performance of the models tested.

#### 4.2.3 Diffusion drift SDE

We present an application from neuroscience, modeling the decision process of a brain. We integrate an SDE that collects information about a particular state of the world:

$$\mathrm{d}x = (v + \lambda x)\mathrm{d}t + \mathrm{d}W, \tag{7}$$

where we differentiate between the constant drift ($\lambda = 0$) and leaky drift ($\lambda < 0$) models. The prior is chosen such that the models are not nested and are shown in Table 6. We integrate until we hit a collapsing symmetric boundary $B(t) = \pm Be^{-\kappa t}$. We record the reaction time of hitting this boundary as the result of the run, with $\pm$ indexing whether the positive or negative boundary was hit. We also add the non-decision time as an additional parameter to the recorded decision time, which accounts for delays in the decision making process. For each parameter combination, we simulate 400 times and record the reaction times. The data vector is the set of those 400 times. In order to respect exchangeability, we use permutation equivariant / invariant neural networks for the flow and the EN / CTSM.

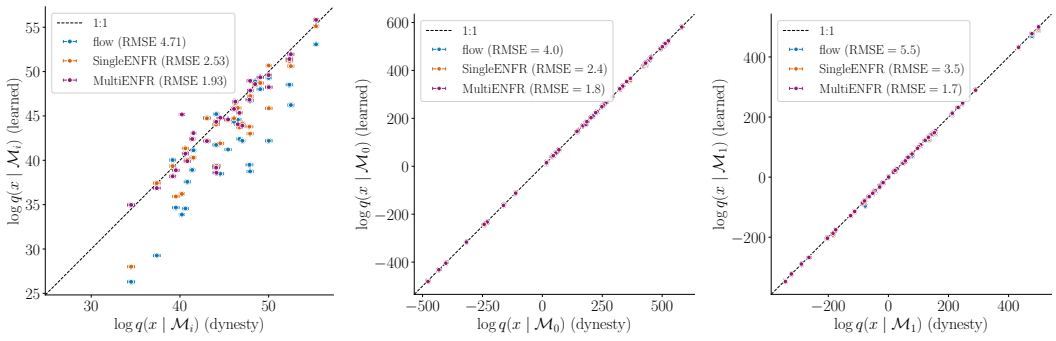
We show our results for the flow, SingleENFR and MultiENFR in Figure 3b. We find that the flow alone can already produce precise estimates of the evidence for both the constant drift and the leaky drift model. Adding the SingleENFR correction however again improves the resulting estimates in RMSE. Our best model is MultiENFR, which produces the highest precision in evidence estimates.

### 4.3 Simulation-based applications

In this section we present examples where the true likelihood is not available, and therefore no likelihood-based estimate of the evidence can be compared to. As we have no way of assessing the evidence estimates in this case, we instead perform blind coverage testing to test the resulting Bayes factors (Zadrozny & Elkan, 2002; Niculescu-Mizil & Caruana, 2005; Jeffrey & Wandelt, 2024). Given the evidence estimate, we compute the model posteriors $p_i = p(\mathcal{M}_i|x)$. For a given model $\mathcal{M}_i$, we can then bin the model posterior predictions for a sample of test simulations in probability bins. Among simulations assigned posterior probability $p_i$ to model $\mathcal{M}_i$, approximately a fraction $p_i$ should actually have been generated from that model. As we know the underlying ground truth model to each simulation we can perform such a test.

#### 4.3.1 The Hodgkin-Huxley model

The Hodgkin-Huxley model is a biophysical model for the behavior of biological neurons. The neuron membrane is modeled as an electrical circuit and the observed quantity is the membrane voltage $V(t)$, whose dynamics are given by an SDE:

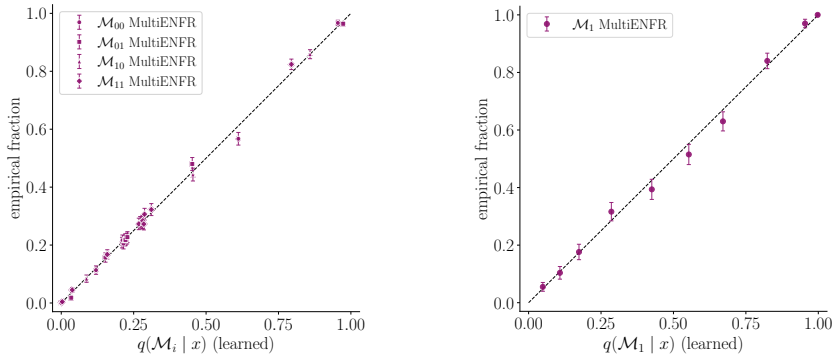



(a) (b)

Figure 3: (a) Evidence estimation for noisy observations of an ODE driven time series using flow matching and our ENFR models. MultiENFR outperforms the other models. (b) Diffusion drift application of standard flow matching and our ENFR networks. MultiENFR shows the best performance and good evidence estimation over a large dynamical range.

$$C_m dV = \big(I_{\text{ext}}(t) - I_{\text{Na}} - I_{\text{K}} - I_{\text{L}} - z_0 I_{\text{M}} - z_1 I_{\text{CaL}}\big)dt + \sigma dW, \tag{8}$$

where $C_m$ is a capacitance, and $I$ are currents induced by present elements. Each $I$ except for $I_L$ and $I_{\text{ext}}$ are described by so-called gating variables which are defined by a system of coupled ODEs driven by the voltage, forming one coupled system. We consider four different models $\mathcal{M}_{z_0 z_1}$, the first of which only contains the sodium (Na), potassium (K) and leak (L) contributions ($\mathcal{M}_{00}$). The other three models are extensions adding the slow potassium (M) ($\mathcal{M}_{10}$), and L-type calcium (CaL) ($\mathcal{M}_{01}$) contributions, or both ($\mathcal{M}_{11}$). We give more details in Appendix A.6. Figure 4a shows the blind calibration test for our MultiENFR model. We find that the resulting Bayes factors are well calibrated.




(a) (b)

Figure 4: Blind coverage testing for the simulation-based applications for (a) the Hodgkin-Huxley application, and (b) the Galaxy assembly bias application. As we do not have access to a likelihood-based evidence estimate for our simulation based models we test whether the resulting Bayes factors are well calibrated. We find that our MultiENFR architecture is well calibrated in both cases. The error bars show $2\sigma$ binomial errors.

### 4.3.2 Galaxy assembly bias

In cosmology, we observe the 3D distribution of galaxies in the Universe as a biased tracer of the underlying dark matter density field. However, theoretical predictions can only accurately model the dark matter distribution. It is therefore of major interest to model the connection between dark matter and galaxies, the so-called bias, and understand what are its major drivers. We compare two

bias models using 362-dimensional summary statistics from simulated galaxy catalogs: a mass-only model and an extension with assembly and phase-space bias. Details and parameter priors are given in Appendix A.7. Figure 4b shows that the MultiENFR model is well calibrated and could be used to determine the underlying bias model from observations in a likelihood-free way.

## 5 Fast evidences in action: hypothesis testing

As all of the models presented here can estimate evidences extremely fast once they are trained, we can use the evidence in traditionally intractable ways. For example, we can turn the evidence into a test statistic and run a hypothesis test (Box, 1980). We conduct such an analysis by drawing 1000 samples from the prior predictive of our baryonic acoustic oscillations application, and use our trained MultiENFR model to estimate the evidence of each simulation under the ΛCDM model. We consider this our null hypothesis. Comparing this to the evidence estimate of the data, allows us to quantify a prior predictive p-value:

$$p^{pp}_{\text{lower}} = P[q_{\phi,\psi}(x_{pp}|\mathcal{M}) \leq q_{\phi,\psi}(x_{\text{obs}}|\mathcal{M})], \tag{9}$$

i.e. we can estimate how many simulations have lower evidence than the data under a given model. We use the finite sample estimator $\hat{p}^{pp}_{\text{lower}} = \frac{R+1}{n+1}$, where $R$ is the number of simulations with lower or equal evidence.

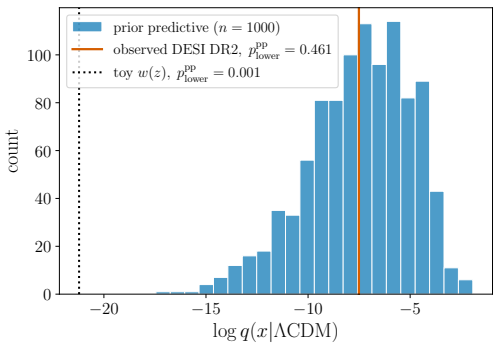



Figure 5: Evidence estimates of the MultiENFR model on prior predictive simulations and the DESI DR2 data under the ΛCDM model. While 46.1% of simulations have lower evidence than the data, no simulation has lower evidence than the toy $w(z)$ model, and we can clearly pick up on the model misspecification.

Figure 5 shows the evidence estimates of 1000 test simulations from the prior predictive as well as the evidence estimate of the DESI DR2 data for the ΛCDM model. We infer $p^{pp}_{\text{lower}} = 0.461$ finding no indication of model misspecification for our dataset. We also show a simulation with a very non-standard dark energy parameterization (see Appendix A.3) and could reject the null hypothesis at $p^{pp}_{\text{lower}} = 0.001$.

## 6 Limitations and Conclusions

Two main limitations remain. First, in our GMM scaling test the flow reference degrades beyond $\sim$ 500 dimensions and the evidence network recovers only a small correction. High-dimensional density estimation remains a bottleneck. Second, our benchmarks use $\mathcal{O}(10^5)$ simulations per model, which many scientific simulators cannot afford. Encouragingly, the evidence-network correction is most effective precisely in the small data regime (Figure 2b), so the method is well placed for the simulation-limited settings where it is needed most.

Within these limits, ENFR provides a practical framework for simulation-based estimation of the Bayesian evidence by combining a flow-matched reference with an evidence network. The single-model version (SingleENFR) improves on the flow it is built on in every benchmark, and the multi-model version (MultiENFR), based on a new pairwise l-POP loss for which we prove asymptotic recovery of all Bayes factors, is the most accurate estimator across our scientific applications. Because inference is amortized, the evidence becomes cheap enough to use as a test statistic, which we demonstrated with a prior predictive misspecification test on DESI DR2 data. Our work pushes the simulation-based paradigm beyond parameter estimation, unlocking more of the standard Bayesian toolkit for the simulation-based case: principled model misspecification testing, goodness-of-fit assessment, and model comparison through the Bayesian evidence.

### AI USE STATEMENT

Generative AI tools were used throughout the preparation of this manuscript. We used Claude Opus 5 and GPT-5.4 and GPT-5.5 for setting up and running numerical experiments and analyses, and discussing the methodology of this project. This includes generation of code which produced the training datasets used in this work. Generated code for our core methods has been checked by the authors. We have also used generative AI to assist in formatting the mathematical proofs presented in this work. GPT-5.5 and GPT-5.6 Sol were used for reviewing this manuscript and stylistic editing. We have not used generative AI tools to draft entire sections of this manuscript. We have not used generative AI tools for the direct generation of any plot. The responsibility stays with the authors.

### ETHICS STATEMENT

We acknowledge the ICLR code of ethics and do not expect our work to have any negative ethical implications. Our work is focused on the evaluation of simulation-based models for scientific reasons.

### REPRODUCIBILITY STATEMENT

We provide the details necessary for reproducing the results presented here. We present in great detail the setups of our experiments along with preprocessing steps in Appendix A. The details about the architectures used are in Appendix B.

### ACKNOWLEDGMENTS

Acknowledgments have been removed for the submission version.

## REFERENCES

M. Abdul Karim, J. Aguilar, S. Ahlen, S. Alam, L. Allen, C. Allende Prieto, O. Alves, A. Anand, U. Andrade, E. Armengaud, A. Aviles, S. Bailey, C. Baltay, P. Bansal, A. Bault, J. Behera, S. Ben-Zvi, D. Bianchi, C. Blake, S. Brieden, A. Brodzeller, D. Brooks, E. Buckley-Geer, E. Burtin, R. Calderon, R. Canning, A. Carnero Rosell, P. Carrilho, L. Casas, F. J. Castander, M. Charles, E. Chaussidon, J. Chaves-Montero, D. Chebat, X. Chen, T. Claybaugh, S. Cole, A. P. Cooper, A. Cuceu, K. S. Dawson, A. de la Macorra, A. de Mattia, N. Deiosso, J. Della Costa, R. Demina, A. Dey, B. Dey, Z. Ding, P. Doel, J. Edelstein, D. J. Eisenstein, W. Elbers, P. Fagrelius, K. Fanning, E. Fernández-García, S. Ferraro, A. Font-Ribera, J. E. Forero-Romero, C. S. Frenk, C. Garcia-Quintero, L. H. Garrison, E. Gaztañaga, H. Gil-Marín, S. Gontcho A Gontcho, D. Gonzalez, A. X. Gonzalez-Morales, C. Gordon, D. Green, G. Gutierrez, J. Guy, B. Hadzhiyska, C. Hahn, S. He, M. Herbold, H. K. Herrera-Alcantar, M.-F. Ho, K. Honscheid, C. Howlett, D. Huterer, M. Ishak, S. Juneau, N. V. Kamble, N. G. Karaçaylı, R. Kehoe, S. Kent, A. G. Kim, D. Kirkby, T. Kisner, S. E. Koposov, A. Kremin, A. Krolewski, O. Lahav, C. Lamman, M. Landriau, D. Lang, J. Lasker, J. M. Le Goff, L. Le Guillou, A. Leauthaud, M. E. Levi, Q. Li, T. S. Li, K. Lodha, M. Lokken, F. Lozano-Rodríguez, C. Magneville, M. Manera, P. Martini, W. L. Matthewson, A. Meisner, J. Mena-Fernández, A. Menegas, T. Mergulhão, R. Miquel, J. Moustakas, A. Muñoz-Gutiérrez, D. Muñoz-Santos, A. D. Myers, S. Nadathur, K. Naidoo, L. Napolitano, J. A. Newman, G. Niz, H. E. Noriega, E. Paillas, N. Palanque-Delabrouille, J. Pan, J. A. Peacock, M. Pellejero Ibanez, W. J. Percival, A. Pérez-Fernández, I. Pérez-Ràfols, M. M. Pieri, C. Poppett, F. Prada, D. Rabinowitz, A. Raichoor, C. Ramírez-Pérez, M. Rashkovetskyi, C. Ravoux, J. Rich, A. Rocher, C. Rockosi, J. Rohlf, J. O. Román-Herrera, A. J. Ross, G. Rossi, R. Ruggeri, V. Ruhlmann-Kleider, L. Samushia, E. Sanchez, N. Sanders, D. Schlegel, M. Schubnell, H. Seo, A. Shafieloo, R. Sharples, J. Silber, F. Sinigaglia, D. Sprayberry, T. Tan, G. Tarlé, P. Taylor, W. Turner, L. A. Ureña-López, R. Vaisakh, F. Valdes, G. Valogiannis, M. Vargas-Magaña, L. Verde, M. Walther, B. A. Weaver, D. H. Weinberg, M. White, M. Wolfson, C. Yèche, J. Yu, E. A. Zaborowski, P. Zarrouk, Z. Zhai, H. Zhang, C. Zhao, G. B. Zhao, R. Zhou, H. Zou, and DESI Collaboration. DESI DR2 results. II. Measurements of baryon acoustic oscillations and cosmological constraints. Phys. Rev. D, 112(8):083515, October 2025. doi: 10.1103/tr6y-kpc6.

S. Agostinelli, J. Allison, K. Amako, J. Apostolakis, H. Araujo, P. Arce, M. Asai, D. Axen, S. Banerjee, G. Barrand, F. Behner, L. Bellagamba, J. Boudreau, L. Broglia, A. Brunengo, H. Burkhardt, S. Chauvie, J. Chuma, R. Chytracek, G. Cooperman, G. Cosmo, P. Degtyarenko, A. Dell'Acqua, G. Depaola, D. Dietrich, R. Enami, A. Feliciello, C. Ferguson, H. Fesefeldt, G. Folger, F. Foppiano, A. Forti, S. Garelli, S. Giani, R. Giannitrapani, D. Gibin, J. J. Gómez Cadenas, I. González, G. Gracia Abril, G. Greeniaus, W. Greiner, V. Grichine, A. Grossheim, S. Guatelli, P. Gumplinger, R. Hamatsu, K. Hashimoto, H. Hasui, A. Heikkinen, A. Howard, V. Ivanchenko, A. Johnson, F. W. Jones, J. Kallenbach, N. Kanaya, M. Kawabata, Y. Kawabata, M. Kawaguti, S. Kelner, P. Kent, A. Kimura, T. Kodama, R. Kokoulin, M. Kossov, H. Kurashige, E. Lamanna, T. Lampén, V. Lara, V. Lefebure, F. Lei, M. Liendl, W. Lockman, F. Longo, S. Magni, M. Maire, E. Medernach, K. Minamimoto, P. Mora de Freitas, Y. Morita, K. Murakami, M. Nagamatu, R. Nartallo, P. Nieminen, T. Nishimura, K. Ohtsubo, M. Okamura, S. O'Neale, Y. Oohata, K. Paech, J. Perl, A. Pfeiffer, M. G. Pia, F. Ranjard, A. Rybin, S. Sadilov, E. Di Salvo, G. Santin, T. Sasaki, N. Savvas, Y. Sawada, S. Scherer, S. Sei, V. Sirotenko, D. Smith, N. Starkov, H. Stoecker, J. Sulkimo, M. Takahata, S. Tanaka, E. Tcherniaev, E. Safai Tehrani, M. Tropeano, P. Truscott, H. Uno, L. Urban, P. Urban, M. Verderi, A. Walkden, W. Wander, H. Weber, J. P. Wellisch, T. Wenaus, D. C. Williams, D. Wright, T. Yamada, H. Yoshida, D. Zschiesche, and G EANT4 Collaboration. GEANT4—a simulation toolkit. *Nuclear Instruments and Methods in Physics Research A*, 506 (3):250–303, July 2003. doi: 10.1016/S0168-9002(03)01368-8.

Shadab Alam et al. The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: cosmological analysis of the DR12 galaxy sample. *Monthly Notices of the Royal Astronomical Society*, 470(3):2617–2652, 2017. doi: 10.1093/mnras/stx721. URL https://arxiv.org/abs/1607.03155.

Justin Alsing, Benjamin Wandelt, and Stephen Feeney. Massive optimal data compression and density estimation for scalable, likelihood-free inference in cosmology. MNRAS, 477(3):2874–2885, July 2018. doi: 10.1093/mnras/sty819.

George E. P. Box. Sampling and bayes' inference in scientific modelling and robustness. *Royal Statistical Society. Journal. Series A: General*, 143(4):383–404, 07 1980. ISSN 0035-9238. doi: 10.2307/2982063. URL https://doi.org/10.2307/2982063.

Johann Brehmer, Gilles Louppe, Juan Pavez, and Kyle Cranmer. Mining gold from implicit models to improve likelihood-free inference. *Proceedings of the National Academy of Sciences*, 117(10): 5242–5249, 2020. ISSN 0027-8424. doi: 10.1073/pnas.1915980117. URL https://www.pnas.org/content/117/10/5242.

Ricky T. Q. Chen. torchdiffeq, 2018. URL https://github.com/rtqichen/torchdiffeq.

Wei Chen, Shigui Li, Jiacheng Li, Junmei Yang, John Paisley, and Delu Zeng. Dequantified diffusion-schrödinger bridge for density ratio estimation. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 8427–8452. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/chen25aj.html.

Kristy Choi, Madeline Liao, and Stefano Ermon. Featurized density ratio estimation. In Cassio de Campos and Marloes H. Maathuis (eds.), *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pp. 172–182. PMLR, 27–30 Jul 2021. URL https://proceedings.mlr.press/v161/choi21a.html.

Kristy Choi, Chenlin Meng, Yang Song, and Stefano Ermon. Density ratio estimation via infinitesimal classification. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 2552–2573. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/choi22a.html.

Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Science*, 117(48):30055–30062, December 2020. doi: 10.1073/pnas.1912789117.

Michael Deistler, Kyra L Kadhim, Matthijs Pals, Jonas Beck, Ziwei Huang, Manuel Gloeckler, Janne K Lappalainen, Cornelius Schröder, Philipp Berens, Pedro J Gonçalves, et al. Jaxley: differentiable simulation enables large-scale training of detailed biophysical models of neural dynamics. *Nature Methods*, pp. 1–9, 2025.

F. Feroz, M. P. Hobson, and M. Bridges. MULTINEST: an efficient and robust Bayesian inference tool for cosmology and particle physics. MNRAS, 398(4):1601–1614, October 2009. doi: 10.1111/j.1365-2966.2009.14548.x.

Ruiqi Gao, Erik Nijkamp, Diederik P. Kingma, Zhen Xu, Andrew M. Dai, and Ying Nian Wu. Flow contrastive estimation of energy-based models. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7515–7525, 2019. URL https://api.semanticscholar.org/CorpusID:208527348.

Hong Guo, Zheng Zheng, Idit Zehavi, Kyle Dawson, Ramin A. Skibba, Jeremy L. Tinker, David H. Weinberg, Martin White, and Donald P. Schneider. Velocity bias from the small-scale clustering of SDSS-III BOSS galaxies. *Monthly Notices of the Royal Astronomical Society*, 446:578–594, 2015. doi: 10.1093/mnras/stu2120. URL https://arxiv.org/abs/1407.4811.

W. J. Handley, M. P. Hobson, and A. N. Lasenby. polychord: nested sampling for cosmology. MNRAS, 450:L61–L65, June 2015. doi: 10.1093/mnrasl/slv047.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2. URL https://doi.org/10.1038/s41586-020-2649-2.

Andrew Hearin, Duncan Campbell, Erik Tollerud, Peter Behroozi, Benedikt Diemer, Nathaniel J. Goldbaum, Elise Jennings, Alexie Leauthaud, Yao-Yuan Mao, Surhud More, John Parejko, Manodeep Sinha, Brigitta Sipőcz, and Andrew Zentner. Forward modeling of large-scale structure: An open-source approach with Halotools. *The Astronomical Journal*, 154:190, 2017. doi: 10.3847/1538-3881/aa859f. URL https://arxiv.org/abs/1606.04106.

Andrew P. Hearin, Andrew R. Zentner, Frank C. van den Bosch, Duncan Campbell, and Erik Tollerud. Introducing decorated HODs: modelling assembly bias in the galaxy–halo connection. *Monthly Notices of the Royal Astronomical Society*, 460:2552–2570, 2016. doi: 10.1093/mnras/stw840. URL https://arxiv.org/abs/1512.03050.

J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3): 90–95, 2007. doi: 10.1109/MCSE.2007.55.

Niall Jeffrey and Benjamin D. Wandelt. Evidence Networks: simple losses for fast, amortized, neural Bayesian model comparison. *Machine Learning: Science and Technology*, 5(1):015008, March 2024. doi: 10.1088/2632-2153/ad1a4d.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. URL https://api.semanticscholar.org/CorpusID:6628106.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Hamed Masnadi-shirazi and Nuno Vasconcelos. On the design of loss functions for classification: theory, robustness to outliers, and savageboost. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou (eds.), *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008. URL https://proceedings.neurips.cc/paper_files/paper/2008/file/f5deaeeae1538fb6c45901d524ee2f98-Paper.pdf.

Andrei Mesinger, Steven Furlanetto, and Renyue Cen. 21CMFAST: a fast, seminumerical simulation of the high-redshift 21-cm signal. MNRAS, 411(2):955–972, February 2011. doi: 10.1111/j.1365-2966.2010.17731.x.

Alexandru Niculescu-Mizil and Rich Caruana. Obtaining calibrated probabilities from boosting. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, UAI'05, pp. 413–420, Arlington, Virginia, USA, 2005. AUAI Press. ISBN 0974903914.

Dily Duan Yi Ong, David Yallup, and Will Handley. The Bayesian view of DESI DR2 with unimpeded: Evidence and tension in a combined analysis with CMB and supernovae across cosmological models. *arXiv e-prints*, art. arXiv:2603.05472, March 2026. doi: 10.48550/arXiv.2603.05472.

George Papamakarios and Iain Murray. Fast $\varepsilon$-free inference of simulation models with bayesian conditional density estimation. *Advances in Neural Information Processing Systems*, pp. 1028–1036, 2016.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Martin Pospischil, Maria Toledo-Rodriguez, Cyril Monier, Zuzanna Piwkowska, Thierry Bal, Yves Frégnac, Henry Markram, and Alain Destexhe. Minimal hodgkin-huxley type models for different classes of cortical and thalamic neurons. *Biological Cybernetics*, 99(4-5):427–441, 2008. doi: 10.1007/s00422-008-0263-8.

Benjamin Rhodes, Kai Xu, and Michael Gutmann. Telescoping density-ratio estimation. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 4905–4916. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/33d3b157ddc0896addfb22fa2a519097-Paper.pdf.

Maxwell Shinn, Norman H Lam, and John D Murray. A flexible framework for simulating and fitting generalized drift-diffusion models. *eLife*, 9:e56938, aug 2020. ISSN 2050-084X. doi: 10.7554/eLife.56938. URL https://doi.org/10.7554/eLife.56938.

Yang Song and Stefano Ermon. *Generative modeling by estimating gradients of the data distribution*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Yang Song, Jascha Sohl-dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Joshua S. Speagle. DYNESTY: a dynamic nested sampling package for estimating Bayesian posteriors and evidences. MNRAS, 493(3):3132–3158, April 2020. doi: 10.1093/mnras/staa278.

A. Thomsen, J. Bucko, T. Kacprzak, V. Ajani, J. Fluri, A. Refregier, D. Anbajagane, F. J. Castander, A. Ferté, M. Gatti, N. Jeffrey, A. Alarcon, A. Amon, K. Bechtol, M. R. Becker, G. M. Bernstein, A. Campos, A. Carnero Rosell, C. Chang, R. Chen, A. Choi, M. Crocce, C. Davis, J. DeRose, S. Dodelson, C. Doux, K. Eckert, J. Elvin-Poole, S. Everett, P. Fosalba, D. Gruen, I. Harrison, K. Herner, E. M. Huff, M. Jarvis, N. Kuropatkin, P.-F. Leget, N. MacCrann, J. McCullough, J. Myles, A. Navarro-Alsina, S. Pandey, A. Porredon, J. Prat, M. Raveri, M. Rodriguez-Monroy, R. P. Rollins, A. Roodman, E. S. Rykoff, C. Sánchez, L. F. Secco, E. Sheldon, T. Shin, M. A.

Troxel, I. Tutusaus, T. N. Varga, N. Weaverdyck, R. H. Wechsler, B. Yanny, B. Yin, Y. Zhang, J. Zuntz, M. Aguena, S. Allam, F. Andrade-Oliveira, D. Bacon, J. Blazek, D. Brooks, R. Camilleri, J. Carretero, R. Cawthon, L. N. da Costa, M. E. da Silva Pereira, T. M. Davis, J. De Vicente, S. Desai, P. Doel, J. García-Bellido, G. Gutierrez, S. R. Hinton, D. L. Hollowood, K. Honscheid, D. J. James, K. Kuehn, O. Lahav, S. Lee, J. L. Marshall, J. Mena-Fernández, F. Menanteau, R. Miquel, J. Muir, R. L. C. Ogando, A. A. Plazas Malagón, E. Sanchez, D. Sanchez Cid, I. Sevilla-Noarbe, M. Smith, E. Suchyta, M. E. C. Swanson, D. Thomas, C. To, D. L. Tucker, and DES Collaboration. Dark Energy Survey Year 3 results: Simulation-based wCDM inference from weak lensing and galaxy clustering maps with deep learning: Analysis design. Phys. Rev. D, 113(8):083501, April 2026. doi: 10.1103/3sj1-1l9f.

Jesús Torrado and Antony Lewis. Cobaya: code for Bayesian analysis of hierarchical physical models. J. Cosmology Astropart. Phys., 2021(5):057, May 2021. doi: 10.1088/1475-7516/2021/05/057.

Francisco Villaescusa-Navarro, Daniel Anglés-Alcázar, Shy Genel, David N. Spergel, Rachel S. Somerville, Romeel Dave, Annalisa Pillepich, Lars Hernquist, Dylan Nelson, Paul Torrey, Desika Narayanan, Yin Li, Oliver Philcox, Valentina La Torre, Ana Maria Delgado, Shirley Ho, Sultan Hassan, Blakesley Burkhart, Digvijay Wadekar, Nicholas Battaglia, Gabriella Contardo, and Greg L. Bryan. The CAMELS Project: Cosmology and Astrophysics with Machine-learning Simulations. ApJ, 915(1):71, July 2021. doi: 10.3847/1538-4357/abf7ba.

Francisco Villaescusa-Navarro et al. The Quijote simulations. *The Astrophysical Journal Supplement Series*, 250:2, 2020. doi: 10.3847/1538-4365/ab9d82. URL https://arxiv.org/abs/1909.05273.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Maximilian von Wietersheim-Kramsta, Kiyam Lin, Nicolas Tessore, Benjamin Joachimi, Arthur Loureiro, Robert Reischke, and Angus H. Wright. KiDS-SBI: Simulation-based inference analysis of KiDS-1000 cosmic shear. A&A, 694:A223, February 2025. doi: 10.1051/0004-6361/202450487.

Hanlin Yu, Arto Klami, Aapo Hyvarinen, Anna Korba, and Omar Chehab. Density ratio estimation with conditional probability paths. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=Gn2izAiYzZ.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 23-26, 2002, Edmonton, Alberta, Canada*, pp. 694–699. ACM, 2002. doi: 10.1145/775047.775151. URL https://doi.org/10.1145/775047.775151.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbhakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J Smola. Deep sets. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 3394–3404, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Zheng Zheng, Alison L. Coil, and Idit Zehavi. Galaxy evolution from halo occupation distribution modeling of DEEP2 and SDSS galaxy clustering. *The Astrophysical Journal*, 667:760–779, 2007. doi: 10.1086/521074. URL https://arxiv.org/abs/astro-ph/0703457.

## A Details about the experiments

Here we provide more details about the numerical experiments we ran.

### A.1 GMM

Our GMM is defined as follows:

$$p(x) = \sum_{k=1}^{3} w_k \, \mathcal{N}(x \mid \mu_k, \Sigma_k) \qquad w = (0.45,\, 0.35,\, 0.20). \tag{10}$$

The components are placed on a sphere of radius $r = 3.8$, i.e. $\mu_k = r\, u_k$ with $u_k$ sampled uniformly from the unit sphere. Each covariance is constructed as

$$\Sigma_k = D_k + \beta\, L_k L_k^{\mathsf{T}} + \epsilon I, \tag{11}$$

where $D_k$ is a diagonal matrix with entries $d_{k,i} \sim \mathcal{U}[0.7, 1.8]$, $L_k \sim \mathcal{N}(0, 1/\sqrt{3}) \in \mathbb{R}^{d\times 3}$ is a random factor, $\beta = 0.55$ controls correlation strength and is fixed in our analysis, and $\epsilon = 0.15$ is a jitter term ensuring positive definiteness and stable inverses.

For the dimensionality scaling we draw 5M samples. 4.5M are used for training and 500k for the evaluation set for the flow. EN and CTSM use 4M for training and 1M for validation for the dimensionality scaling. We test on 200 more fresh draws from the GMM, which are not used in training or evaluation. We preprocess the data by standardizing it with parameters estimated from the training set:

$$x = \frac{x_{\text{raw}} - \mu_{\text{train}}}{\sigma_{\text{train}}}, \tag{12}$$

where $\mu_{\text{train}}$ is the estimated mean of the training set and $\sigma_{\text{train}}$ is the estimated standard deviation of the training set. As we end up estimating densities, we have to take into account the Jacobian $J$ of this transformation:

$$\log p(x_{\text{raw}}) = \log p(x) + \log \left|\det J\right|. \tag{13}$$

The entire runtime for the dimensionality scaling of all models for one of the seeds presented here was $\sim 55$ hours in wall clock, where we trained on a single NVIDIA A40 48GB VRAM.

### A.2 BAO AND PRIORS

Baryonic acoustic oscillations are a characteristic scale in the Universe that is set by traveling sound waves in the early Universe. At some point, due to temperature changes, these sound waves can no longer propagate and are frozen in. This imprints the sound horizon scale which is the distance of how far a sound wave could have traveled until it is frozen. Today, we can still observe this scale by measuring the statistics of the galaxies that have formed.

The observed BAO scale depends on the expansion history of the Universe. Along the line of sight, it is related to the Hubble distance,

$$D_H(z) = \frac{c}{H(z)}, \tag{14}$$

while perpendicular to the line of sight it is related to the transverse comoving distance,

$$D_M(z) = \begin{cases} \dfrac{c}{H_0\sqrt{\Omega_k}} \sinh\left[\sqrt{\Omega_k}\, \dfrac{H_0 D_C(z)}{c}\right], & \Omega_k > 0, \\ D_C(z), & \Omega_k = 0, \\ \dfrac{c}{H_0\sqrt{|\Omega_k|}} \sin\left[\sqrt{|\Omega_k|}\, \dfrac{H_0 D_C(z)}{c}\right], & \Omega_k < 0, \end{cases} \tag{15}$$

Table 3: Boundaries of the uniform priors for the parameters of the five cosmological models used in our BAO example. The six base parameters of the ΛCDM model are shared by every model. Each extension introduces either one or two more parameters.

| Parameter | ΛCDM | $w$CDM | $w_0w_a$CDM | $\nu$CDM | $\Omega_k$CDM |
|---|---|---|---|---|---|
| Base parameters (shared) | | | | | |
| $H_0$ | | | $[20, 100]$ | | |
| $\Omega_b h^2$ | | | $[0.005, 0.1]$ | | |
| $\Omega_c h^2$ | | | $[0.001, 0.99]$ | | |
| $\tau$ | | | $[0.01, 0.8]$ | | |
| $\log(10^{10} A_s)$ | | | $[1.61, 3.91]$ | | |
| $n_s$ | | | $[0.8, 1.2]$ | | |
| Extension parameters | | | | | |
| $w$ (or $w_0$) | — | $[-3, -0.333]$ | $[-3, 1]$ | — | — |
| $w_a$ | — | — | $[-3, 2]$ | — | — |
| $\sum m_\nu$ [eV] | — | — | — | $[0.06, 2]$ | — |
| $\Omega_k$ | — | — | — | — | $[-0.3, 0.3]$ |
| Number of Parameters | 6 | 7 | 8 | 7 | 7 |

where

$$D_C(z) = c \int_0^z \frac{\mathrm{d}z'}{H(z')}. \tag{16}$$

Here, $c$ is the speed of light, $z$ is the redshift, and $H(z)$ is the Hubble expansion rate at redshift $z$. When the BAO signal is measured isotropically, these two distances are combined into the volume-averaged distance,

$$D_V(z) = \left[ z\, D_H(z)\, D_M^2(z) \right]^{1/3}. \tag{17}$$

BAO measurements therefore constrain the dimensionless combinations $D_H(z)/r_d$, $D_M(z)/r_d$, or $D_V(z)/r_d$, depending on the analysis, where $r_d$ is the sound horizon at the baryon drag epoch. The data vector consists of 13 of these quantities measured at 7 different redshifts.

Table 3 shows the boundaries of our uniform priors on the parameters of the cosmological models. We follow the choices by Ong et al. (2026). We run simulations using the Gaussian likelihood and the DESI DR2 covariance Abdul Karim et al. (2025) using the `cobaya` (Torrado & Lewis, 2021) code. We run 500k simulations of each model for training and 2k for evaluation. For each model we run nested sampling for an additional 100 simulations per model. Due to the presence of outliers in the data, before applying Equation 12, we apply the transform:

$$x = \operatorname{asinh}\left( \frac{x_{\mathrm{raw}} - \mathrm{median}(x_{\mathrm{raw}})}{\mathrm{IQR}} \right), \tag{18}$$

where IQR is the range between the 25th and 75th percentile. We subsequently standardize according to Equation 12. For density estimation, the Jacobian of this transformation is taken into account in our experiments.

We ran the BAO experiments on a single NVIDIA A40 with 48GB VRAM.

### A.3 Toy dark energy model

As a toy model for late-time model misspecification, we consider a smooth transition in the dark-energy equation of state,

$$w(z) = -1 + \frac{\Delta w}{2} \left[ 1 - \tanh\left( \frac{z - z_t}{\Delta z} \right) \right]. \tag{19}$$

Table 4: Prior distributions for the ODE oscillator models. All priors are independent uniforms; parameters shared between models have identical priors.

| Parameter | Prior | Harmonic | Damped | Duffing |
|---|---|---|---|---|
| $x_0$ | $\mathcal{U}(-2,\, 2)$ | ✓ | ✓ | ✓ |
| $v_0$ | $\mathcal{U}(-2,\, 2)$ | ✓ | ✓ | ✓ |
| $\omega$ | $\mathcal{U}(0.4,\, 2.5)$ | ✓ | ✓ | ✓ |
| $\gamma$ | $\mathcal{U}(0,\, 0.4)$ | | ✓ | ✓ |
| $\beta$ | $\mathcal{U}(-0.8,\, 0.8)$ | | | ✓ |

Table 5: RMSE of the learned evidence against nested sampling (dynesty) on the ODE oscillator benchmark, over 10 held-out simulations per model.

| Method | harmonic | damped | Duffing | All |
|---|---|---|---|---|
| flow | 6.13 | 3.69 | 3.91 | 4.71 |
| SingleENFR | 2.54 | 2.44 | 2.60 | 2.53 |
| MultiENFR | 2.36 | 0.94 | 2.17 | 1.93 |
| flow + CTSM-SB (ens.) | 135.47 | 26.23 | 7.96 | 79.80 |

We choose

$$\Delta w = 0.4, \qquad z_t = 0.6, \qquad \Delta z = 0.15. \tag{20}$$

The model therefore approaches $w = -1$ at high redshift, while at low redshift it transitions to $w \simeq -0.6$.

### A.4 ODE APPLICATION DETAILS AND PRIORS

Table 4 shows the priors on the parameters. We generate a training set of 1M simulations per model, 10% of which is used as validation data. We apply the preprocessing transformation from Equation 12. We quote our results in Table 5.

### A.5 DDM

We show the priors of our DDM experiment in Table 6. The likelihood is available by solving the corresponding Fokker-Planck equation. We use the pyDDM implementation for this and refer to its documentation for further details (Shinn et al., 2020). We generate 500k simulations for each model for training out of which we split 10% for validation. We apply the preprocessing transformation from Equation 18. We show our results in Table 7.

### A.6 HODGKIN-HUXLEY

The currents are defined as follows:

$$\begin{aligned} I_{\mathrm{Na}} &= g_{\mathrm{Na}}, m^3 h, (V - E_{\mathrm{Na}}) \\ I_{\mathrm{K}} &= g_{\mathrm{K}} n^4 (V - E_{\mathrm{K}}) \\ I_{\mathrm{L}} &= g_L (V - E_L) \\ I_{\mathrm{M}} &= g_M p (V - E_{\mathrm{K}}) \\ I_{\mathrm{CaL}} &= g_{\mathrm{Ca}}, q^2 r, (V - E_{\mathrm{Ca}}), \end{aligned} \tag{21}$$

where $g$ are the maximal conductances, $E$ are reversal potentials and $m, h, n, p, q, r \in [0, 1]$ are gating variables. The gating variables $x_{\mathrm{gate},a} \in \{m, h, n\}$ are obtained by a first order differential equation: $\dot{x}_{\mathrm{gate},a} = \alpha_x(V - V_t)(1 - x_{\mathrm{gate},a}) - \beta_x(V - V_t)\, x_{\mathrm{gate},a}$, where $\alpha$ and $\beta$ are voltage-dependent rate functions as defined in Pospischil et al. (2008). The gating variables $x_{\mathrm{gate},b} \in \{q, r\}$ evolve according to $\dot{x}_{\mathrm{gate}_b} = \alpha_x(V)(1 - x_{\mathrm{gate}_b}) - \beta_x(V)\, x_{\mathrm{gate},b}$. p is obtained through $\dot{p} = (p_\infty(V) - p)/\tau_p(V)$ where $p_\infty$ and $\tau_p$ are defined by:

Table 6: Prior distributions for the drift-diffusion models.

| Parameter | Description | Prior | $\mathcal{M}_0$ | $\mathcal{M}_1$ |
|---|---|---|---|---|
| $v$ | drift rate | $\mathcal{U}(0,\,5)$ | ✓ | ✓ |
| $\lambda$ | leak rate | $\mathcal{U}(-20,\,-5)$ | | ✓ |
| $B$ | initial boundary height | $\mathcal{U}(0.3,\,2)$ | ✓ | ✓ |
| $\kappa$ | boundary collapse rate | $\mathcal{U}(0.5,\,1.5)$ | ✓ | ✓ |
| $t_{\mathrm{nd}}$ | non-decision time [s] | $\mathcal{U}(0.1,\,0.3)$ | ✓ | ✓ |

Table 7: Estimated evidence against nested sampling (dynesty) on the DDM benchmark, over 50 held-out datasets from each model.

| Model | RMSE under $\mathcal{M}_0$ | RMSE under $\mathcal{M}_1$ |
|---|---|---|
| flow | 4.0 | 5.5 |
| SingleENFR | 2.4 | 3.5 |
| MultiENFR | 1.8 | 1.7 |
| flow + CTSM-SB | 20.3 | 115.4 |

$$
\begin{aligned}
p_\infty &= 1/(1+e^{-(V+35)/10}) \\
\tau_p &= \tau_M/(3.3e^{(V+35)/20}+e^{-(V+35)/20}).
\end{aligned}
\tag{22}
$$

All gates are initialized at $V(0) = -70\,\mathrm{mV}$. The injected current is a step of $2\,\mu\mathrm{A\,cm^{-2}}$ between 225 and 1225 ms with added noise, $I_{\mathrm{inj}}(t_k) = I_{\mathrm{step}}(t_k) + \sigma\,\xi_k/\sqrt{\Delta t}$ where $\xi_k \sim \mathcal{N}(0,1)$. We integrate the voltage using a Crank–Nicolson integrator at $\Delta t = 0.025$ ms.

From each simulation we extract 13 hand-picked summaries: resting potential, spike latency, inverse first/mean/median inter-spike interval, two ISI-adaptation indices, two AP-amplitude adaptation indices, AP amplitude, inverse AP width, and the coefficient of variation and Fano factor of the AP amplitudes. Some summaries are undefined in some of the simulations, e.g. due to the absence of spikes. In that case they are set to their physical limit (latency 1000 ms, inverse ISI and inverse width 0, adaptation indices 1, amplitude, CV and Fano factor 0).

We use the `jaxley` Deistler et al. (2025) library for implementation. We generate 400k simulations of each model for training and use 10% of that for validation. We present our prior choices in Table 8 and list fixed parameters in Table 9. We preprocess using Equation 12.

### A.7 Galaxy assembly bias

We populate halo catalogs from 2000 high-resolution Quijote Latin-hypercube simulations (Villaescusa-Navarro et al., 2020) at $z = 0.5$ in periodic cubes of side length $1000\,h^{-1}\mathrm{Mpc}$. Model $\mathcal{M}_0$ uses a mass-only halo occupation distribution (HOD) (Zheng et al., 2007), implemented with Halotools (Hearin et al., 2017), and baseline mean occupations

$$
\begin{aligned}
\mu_{\mathrm{cen}}(M) &= \frac{1}{2}\left[1+\mathrm{erf}\left(\frac{\log_{10}M-\log_{10}M_{\mathrm{min}}}{\sigma_{\log M}}\right)\right], \\
\mu_{\mathrm{sat}}(M) &= \mu_{\mathrm{cen}}(M)\left(\frac{M-M_0}{M_1}\right)^{\alpha}, \qquad M > M_0,
\end{aligned}
\tag{23}
$$

with $\mu_{\mathrm{sat}} = 0$ otherwise. Central and satellite counts are drawn separately from Bernoulli and Poisson distributions. Centrals follow the halo centers and bulk velocities while satellites have Navarro–Frenk–White radial profiles and Jeans-equilibrium velocities. Model $\mathcal{M}_1$ adds concentration-dependent occupation at fixed mass using a decorated HOD (Hearin et al., 2016), with strengths $A_{\mathrm{cen}}$ and $A_{\mathrm{sat}}$. It also includes velocity bias (e.g. Guo et al., 2015) and a rescaled satellite concentration: central velocity dispersion $\alpha_{\mathrm{c}}V_{\mathrm{vir}}$, satellite velocity multiplier $\alpha_{\mathrm{s}}$ relative to the halo,

Table 8: Prior distributions for the Hodgkin–Huxley models.

| Parameter | Prior | Unit |
|---|---|---|
| $g_L$ | $\mathcal{U}(10^{-6},\, 3\times10^{-4})$ | $\mathrm{S\,cm^{-2}}$ |
| $E_L$ | $\mathcal{U}(-80,\, -60)$ | $\mathrm{mV}$ |
| $g_\mathrm{K}$ | $\mathcal{U}(1.5\times10^{-3},\, 1.5\times10^{-2})$ | $\mathrm{S\,cm^{-2}}$ |
| $V_t$ | $\mathcal{U}(-70,\, -50)$ | $\mathrm{mV}$ |
| $g_\mathrm{Na}$ | $\mathcal{U}(8\times10^{-3},\, 8\times10^{-2})$ | $\mathrm{S\,cm^{-2}}$ |
| $\sigma$ | $\mathcal{U}(10^{-4},\, 0.15)$ | $\mu\mathrm{A\,cm^{-2}\,ms^{1/2}}$ |
| $g_\mathrm{M}$ | $\mathcal{U}(10^{-5},\, 6\times10^{-4})$ | $\mathrm{S\,cm^{-2}}$ |
| $\tau_\mathrm{M}$ | $\mathcal{U}(200,\, 2000)$ | $\mathrm{ms}$ |
| $g_\mathrm{Ca}$ | $\mathcal{U}(5\times10^{-5},\, 10^{-3})$ | $\mathrm{S\,cm^{-2}}$ |

Table 9: Fixed parameters for the Hodgkin–Huxley simulations.

| Quantity | Value | Unit |
|---|---|---|
| $E_\mathrm{Na}$ | 50 | $\mathrm{mV}$ |
| $E_\mathrm{K}$ | $-90$ | $\mathrm{mV}$ |
| $E_\mathrm{Ca}$ | 120 | $\mathrm{mV}$ |
| $C_m$ | 1 | $\mu\mathrm{F\,cm^{-2}}$ |
| Compartment length / radius | 10 / 1 | $\mu\mathrm{m}$ |
| Step current density (225–1225 ms) | 2 | $\mu\mathrm{A\,cm^{-2}}$ |
| Trace duration / integration step | 1450 / 0.025 | $\mathrm{ms}$ |

and satellite concentration $c_\mathrm{gal} = F_\mathrm{sat} c_\mathrm{halo}$. For more details on the HOD models see Zheng et al. (2007); Hearin et al. (2016).

Table 10 lists the priors in our analysis. Shared parameters have identical priors and the models have equal prior probability. For $\log_{10} M_\mathrm{min}$, we uniformly sample the $z = 0.5$ entries of $10^6$ stored draws from an empirical prior separately constrained using the redshift distribution of the BOSS DR12 CMASS North Galactic Cap sample (Alam et al., 2017). Other HOD parameters are drawn independently as tabulated. No HOD redshift evolution is used.

The LHID-disjoint dataset split contains 64,000 training simulations (32,000 per model), 8,000 validation simulations and 8,000 frozen test simulations; the estimators also reserve internal validation subsets from the training partition. A single flow-matching model is fitted to the pooled training set, and 32,000 generated samples form a third class alongside the two equally sized real-model classes. We train a five-member ensemble of three-output multiclass Evidence Networks with the pairwise loss, using standardization from Equation 12 for both the flow and the networks.

## B Architectures

We present the architectures of the neural networks used in this work. All networks are implemented in `PyTorch` (Paszke et al., 2019). We have also made use of the `numpy` (Harris et al., 2020), `scipy` (Virtanen et al., 2020), `sklearn` (Pedregosa et al., 2011), `torchdiffeq` (Chen, 2018), and `matplotlib` (Hunter, 2007) libraries.

**Software and data licenses** PyTorch, NumPy, SciPy, and scikit-learn are used under their BSD-3-Clause licenses; dynesty, pyddm, and torchdiffeq under the MIT license; Cobaya under LGPL-3.0-or-later with its stated additional conditions; jaxley under the Apache 2.0 licence; and Matplotlib under the Matplotlib license. The DESI data products are used under CC BY 4.0.

Table 10: Priors for the Quijote galaxy-bias comparison. $\mathcal{U}(a, b)$ denotes a uniform distribution and $\mathcal{TN}(0, 0.2^2; [-1, 1])$ a zero-mean Gaussian of standard deviation 0.2, truncated to $[-1, 1]$. Mass logarithms are base ten with masses in $h^{-1}M_\odot$. Numerical constants denote fixed values. The HOD mass $M_1$ is distinct from the model label $\mathcal{M}_1$.

| Parameter | $\mathcal{M}_0$ | $\mathcal{M}_1$ |
|---|---|---|
| $\Omega_\mathrm{m}$ | $\mathcal{U}(0.1, 0.5)$ | $\mathcal{U}(0.1, 0.5)$ |
| $\Omega_\mathrm{b}$ | $\mathcal{U}(0.03, 0.07)$ | $\mathcal{U}(0.03, 0.07)$ |
| $h$ | $\mathcal{U}(0.5, 0.9)$ | $\mathcal{U}(0.5, 0.9)$ |
| $n_\mathrm{s}$ | $\mathcal{U}(0.8, 1.2)$ | $\mathcal{U}(0.8, 1.2)$ |
| $\sigma_8$ | $\mathcal{U}(0.6, 1.0)$ | $\mathcal{U}(0.6, 1.0)$ |
| $\log_{10} M_\mathrm{min}$ | NGC empirical prior | NGC empirical prior |
| $\sigma_{\log M}$ | $\mathcal{U}(0.1, 0.6)$ | $\mathcal{U}(0.1, 0.6)$ |
| $\log_{10} M_0$ | $\mathcal{U}(13, 15)$ | $\mathcal{U}(13, 15)$ |
| $\log_{10} M_1$ | $\mathcal{U}(13, 15)$ | $\mathcal{U}(13, 15)$ |
| $\alpha$ | $\mathcal{U}(0, 1.5)$ | $\mathcal{U}(0, 1.5)$ |
| $A_\mathrm{cen}$ | 0 | $\mathcal{TN}(0, 0.2^2; [-1, 1])$ |
| $A_\mathrm{sat}$ | 0 | $\mathcal{TN}(0, 0.2^2; [-1, 1])$ |
| $\alpha_\mathrm{c}$ | 0 | $\mathcal{U}(0, 0.7)$ |
| $\alpha_\mathrm{s}$ | 1 | $\mathcal{U}(0.2, 2.0)$ |
| $F_\mathrm{sat}$ | 1 | $\mathcal{U}(0.2, 2.0)$ |

### B.1 Flow matching

We parameterize the velocity field of the flow matching algorithm with a 256 unit, 4 hidden layer fully connected multilayer perceptron using SiLU activation functions. We train with uniform time sampling, using the `Adam` Kingma & Ba (2014) optimizer with a batch size of 256 for the Hodgkin and DDM applications and 128 for all other applications. For the GMM and ODE models we use a learning rate of $10^{-4}$, for the BAO, DDM and Hodgkin models $10^{-3}$. For density estimation via integration of the flow, we use the `dopri5` integrator implemented in `torchdiffeq` to solve

$$\log q_\phi(x|\mathcal{M}) = \log p(z) - \int_0^1 \nabla \cdot v_\phi(x_t, t)\, dt, \tag{24}$$

where $x_t$ follows the learned ODE trajectory. We leave the impact of a thorough hyperparameter search and of a different optimizer to further study.

For the DDM application we use permutation equivariant DeepSets (Zaheer et al., 2017). First we apply a neural network to each dimension:

$$h_i = \phi(x_i, t), \tag{25}$$

to the outputs of which we apply permutation invariant aggregation functions:

$$A = \big[\, \overline{h},\ \max_i h_i,\ \min_i h_i,\ \mathrm{std}_i(h_i) \,\big]. \tag{26}$$

The velocity field is then learned from a concatenation of these aggregated features with the individual observation:

$$v_i = \rho\big(x_i,\ h_i,\ A,\ t\big). \tag{27}$$

The neural networks $\phi$ and $\rho$ are parameterized as MLPs with 256 neurons in each of the 3 layers. The output dimension of $\phi$ is also 256.

### B.2 ENFR

The evidence networks are fully connected neural networks. For the GMM example we use a rather large network with 1024 x 4 neurons with the SiLU activation function. For the BAO example we use a substantially smaller network with 128x2 neurons for MultiENFR and 128x3 neurons for SingleENFR with the ReLU activation function. For the ODE application we use 256x4 neurons with ReLU activation. For the Hodgkin application we use 128x2 neurons with ReLU activation. In all examples except for the GMM we train an ensemble of 5 evidence networks with equal weight, while the GMM test only shows single networks. Batch sizes are 128 (GMM,BAO), 256 (Hodgkin, DDM), and 1024 (ODE).

We train using the `Adam` Kingma & Ba (2014) optimizer with a learning rate of $10^{-4}$.

For the DDM experiment MultiENFR is a permutation invariant DeepSet (Zaheer et al., 2017). A first neural network delivers

$$h_i = \phi(x_i), \tag{28}$$

where $x_i$ are the simulated signed decision times. The output is then aggregated:

$$A = \left[\, \overline{h},\ \max_i h_i,\ \min_i h_i,\ \mathrm{std}_i(h_i) \,\right], \tag{29}$$

and fed into a second network

$$y_j = \rho(A). \tag{30}$$

$\rho$ and $\phi$ are parameterized with MLPs with two and three layers respectively, with 128 neurons in each layer, using a ReLU activation function. We train an equally weighted 5 member ensemble with learning rate $10^{-4}$ batch size 256 and patience 10 in early stopping.

### B.3 CTSM

The CTSM models are also fully connected multilayer perceptrons. There are two different interpolants we use, the variance preserving bridge (VP):

$$x_t = \sqrt{1-t^2}z + tx_1 \quad \text{with } z \sim \mathcal{N}(0,1), \tag{31}$$

and the Schrödinger bridge (SB):

$$x_t = (1-t)x_0 + tx_1 + \sigma\sqrt{t(1-t)}\epsilon \quad \text{with } \epsilon \sim \mathcal{N}(0,1), \tag{32}$$

where $x_0 \sim q_\phi(x)$ is drawn from the flow reference and $x_1 \sim p(x|\mathcal{M})$ are data samples. We choose the noise scale $\sigma$ to be 1.

In all CTSM runs, the architectures have always been identical to the corresponding SingleENFR runs.

## C Results of the GMM scaling experiment

We show the results of the GMM scaling in Table 11. We further show an example plot of the improvement that comes from adding the evidence network in 128 dimensions in Figure 6.

## D Bayes factor analysis for DESI DR2 data

Table 12 shows the explicit comparison of our MultiENFR Bayes factors compared to the likelihood-based analysis.

Table 11: RMSE values of the estimated density $\log q(x)$ with respect to the true density $\log p(x)$ over 200 held out test draws (lower is better) averaged over the three seeds. The SingleENFR combination consistently outperforms the other models. The flow is shared at each seed.

| Example | Flow | CTSM-VP | SingleENFR | CTSM-SB + FlowRef |
|---|---|---|---|---|
| GMM 64D | 0.76 | 2.14 | **0.45** | 1.06 |
| GMM 128D | 2.30 | 7.70 | **1.32** | 2.46 |
| GMM 256D | 6.61 | 27.87 | **3.04** | 6.57 |
| GMM 512D | 15.79 | 65.36 | **14.44** | 15.87 |
| GMM 1024D | 34.99 | 163.21 | **33.44** | 35.01 |

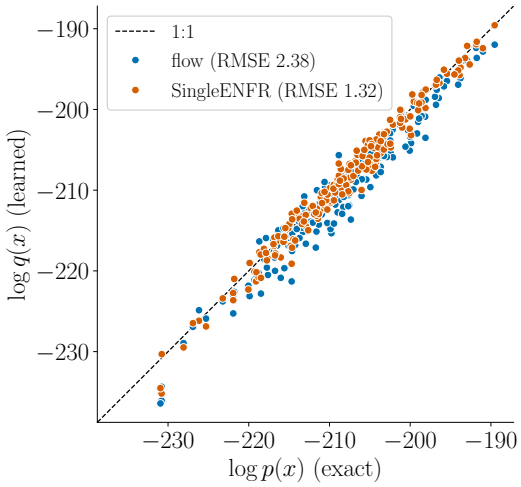



Figure 6: The performance on test points in 128 dimensions for the flow and SingleENFR.

Table 12: Bayes factors w.r.t. ΛCDM on the DESI DR2 data. The error of our simulation-based estimate is largely comparable to the error of the likelihood-based estimate in three of the four models.

| Model | Likelihood Based Analysis | MultiENFR | Δ |
|---|---|---|---|
| $w$CDM | -2.42±0.24 | -1.94 | +0.48 |
| $w_0w_a$CDM | -1.64±0.24 | -1.40 | +0.25 |
| $\nu$CDM | -0.07±0.22 | 0.06 | +0.13 |
| $\Omega_k$CDM | -1.89±0.23 | -1.65 | +0.24 |

## E PROOF OF MINIMIZER OF EVIDENCE NETWORKS

We present the proof that the minimizer of the risk of the l-POP loss is indeed the Bayes factor if the models are weighted equally a priori.

**Proposition.** For a sample $(x, m)$ with $m \in \{0, 1\}$, the l-POP exponential loss is

$$\mathcal{L}(x, m) = \exp\left[\left(\frac{1}{2} - m\right) u(x)\right],$$

using the shorthand notation $u(x) = \mathcal{J}_\alpha(f_\psi(x))$. Assuming equal model sampling probabilities, the population optimum satisfies

$$u^*(x) = \log \frac{p(x \mid \mathcal{M})}{p(x \mid \mathcal{M}_0)} = \log K,$$

with the $^*$ superscript indicating at minimizing parameters $\psi = \hat{\psi}$.

**Proof.** Consider the population risk as a functional of $u(\cdot)$. Under equal sampling probabilities for the two models,

$$P(\mathcal{M}_0) = P(\mathcal{M}) = \frac{1}{2},$$

the expected loss is

$$\mathcal{R}[u] = \frac{1}{2} \int \left[p(x \mid \mathcal{M}_0) e^{u(x)/2} + p(x \mid \mathcal{M}) e^{-u(x)/2}\right] dx.$$

We proceed with standard variational calculus, differentiating with respect to some perturbation $\varepsilon$ at $\varepsilon = 0$:

$$\delta\mathcal{R}[u; \eta] = \frac{d}{d\varepsilon} \mathcal{R}[u + \varepsilon\eta]\Big|_{\varepsilon=0}.$$

Thus,

$$\delta\mathcal{R}[u; \eta] = \frac{1}{4} \int \left[p(x \mid \mathcal{M}_0) e^{u(x)/2} - p(x \mid \mathcal{M}) e^{-u(x)/2}\right] \eta(x)\, dx.$$

At a minimizer $u^*$, the variation must vanish so we find

$$p(x \mid \mathcal{M}_0) e^{u^*(x)/2} - p(x \mid \mathcal{M}) e^{-u^*(x)/2} = 0$$

Hence,

$$p(x \mid \mathcal{M}_0) e^{u^*(x)/2} = p(x \mid \mathcal{M}) e^{-u^*(x)/2}.$$

Multiplying by $e^{u^*(x)/2}$ gives

$$p(x \mid \mathcal{M}_0) e^{u^*(x)} = p(x \mid \mathcal{M}),$$

and therefore

$$e^{u^*(x)} = \frac{p(x \mid \mathcal{M})}{p(x \mid \mathcal{M}_0)}.$$

Taking logarithms,

$$u^*(x) = \log \frac{p(x \mid \mathcal{M})}{p(x \mid \mathcal{M}_0)} = \log K.$$

Thus,

$$\boxed{u^*(x) = \log K}.$$

Without the shorthand this delivers

$$\boxed{\mathcal{J}_\alpha(f^*(x)) = \log K}.$$

## F EN CORRECTION TO THE FLOW UNDER CONSTANT PREDICTION

We use the l-POP loss EN to correct the density estimates of a generative flow. It is interesting to examine the case where the EN cannot learn anything from the data and just returns a minimizing constant, i.e. $f(x) = \text{const}$. In this case we can show, that the EN correction asymptotically does not degrade the flow's evidence estimate under equal model weights.

Consider again the population risk of the l-POP loss of a constant predictor $u(x) = \mathcal{J}_\alpha(f_\psi(x)) = c$ under model priors which are reflected by the training class probabilities $p(\mathcal{M})$ and $p(\mathcal{M}_0)$:

$$R[c] = \int p(\mathcal{M}_0)p(x \mid \mathcal{M}_0)e^{c/2} + p(\mathcal{M})p(x \mid \mathcal{M})e^{-c/2}\, dx = p(\mathcal{M}_0)e^{c/2} + p(\mathcal{M})e^{-c/2}.$$

We consider the variation

$$\delta\mathcal{R}[c; a] = \frac{d}{d\varepsilon}\mathcal{R}[c + \varepsilon a]\bigg|_{\epsilon=0},$$

which delivers

$$\delta\mathcal{R}[c; a] = \frac{a}{2}\left[p(\mathcal{M}_0)e^{c/2} - p(\mathcal{M})e^{-c/2}\right].$$

At the variational optimum $c^*$, the variation must vanish so

$$p(\mathcal{M}_0)e^{c^*/2} - p(\mathcal{M})e^{-c^*/2} = 0,$$

which implies

$$e^{c^*} = \frac{p(\mathcal{M})}{p(\mathcal{M}_0)}.$$

Hence the optimal constant correction is

$$\boxed{c^* = \log \frac{p(\mathcal{M})}{p(\mathcal{M}_0)}}.$$

For equal model weights,

$$p(\mathcal{M}_0) = p(\mathcal{M}),$$

and therefore

$$\boxed{c^* = 0}.$$

The asymptotically optimal EN therefore does not degrade a well-trained flow in this case.

## G Proof of minima of the Multiclass Evidence Network

We present the proof that the multiclass evidence network loss function is minimized by the Bayes factors.

**Proposition.** Consider $N$ models

$$\mathcal{M}_1, \ldots, \mathcal{M}_N$$

and define

$$u_i(x) = \mathcal{J}_\alpha(f_i(x)).$$

For a sample $x$ generated from model $\mathcal{M}_m$, consider the pairwise loss

$$\mathcal{L}(x, m) = \frac{1}{N-1} \sum_{j \neq m} \exp\left[-\frac{1}{2}\left(u_m(x) - u_j(x)\right)\right].$$

Assume that the models are sampled with equal probability,

$$p(\mathcal{M}_i) = \frac{1}{N}.$$

Then the population-risk functional, up to the positive constant factor $1/[N(N-1)]$, is

$$\mathcal{R}[\mathbf{u}] = \int \sum_{i=1}^{N} p(x \mid \mathcal{M}_i) \sum_{j \neq i} \exp\left[-\frac{1}{2}\left(u_i(x) - u_j(x)\right)\right] dx.$$

Define

$$p_i(x) \equiv p(x \mid \mathcal{M}_i).$$

Adding the diagonal terms $i = j$, which contribute only

$$\sum_i p_i(x)$$

and are independent of $\mathbf{u}$, we may equivalently minimize

$$\widetilde{\mathcal{R}}[\mathbf{u}] = \int \left(\sum_i p_i(x) e^{-u_i(x)/2}\right)\left(\sum_j e^{u_j(x)/2}\right) dx.$$

For compactness, define

$$A(x) = \sum_i p_i(x) e^{-u_i(x)/2}, \qquad B(x) = \sum_i e^{u_i(x)/2}.$$

Then

$$\widetilde{\mathcal{R}}[\mathbf{u}] = \int A(x)B(x)\,dx.$$

Again we follow variation of the functional:

$$\delta\widetilde{\mathcal{R}}[\mathbf{u};\boldsymbol{\eta}] = \frac{d}{d\varepsilon}\widetilde{\mathcal{R}}[\mathbf{u}+\varepsilon\boldsymbol{\eta}]\Big|_{\varepsilon=0}.$$

Using the product rule, we first apply this to each term separately:

$$A_\varepsilon(x) = \sum_i p_i(x)e^{-[u_i(x)+\varepsilon\eta_i(x)]/2},$$

and therefore

$$\frac{dA_\varepsilon}{d\varepsilon}\Big|_{\varepsilon=0} = -\frac{1}{2}\sum_i p_i(x)e^{-u_i(x)/2}\eta_i(x).$$

Similarly for the second term,

$$B_\varepsilon(x) = \sum_i e^{[u_i(x)+\varepsilon\eta_i(x)]/2},$$

so that

$$\frac{dB_\varepsilon}{d\varepsilon}\Big|_{\varepsilon=0} = \frac{1}{2}\sum_i e^{u_i(x)/2}\eta_i(x).$$

Now according to the product rule we can write

$$\begin{aligned}\delta\widetilde{\mathcal{R}} &= \int \left[\delta A(x)\,B(x) + A(x)\,\delta B(x)\right]dx \\ &= \frac{1}{2}\int \sum_{k=1}^{N}\left[A(x)e^{u_k(x)/2} - B(x)p_k(x)e^{-u_k(x)/2}\right]\eta_k(x)\,dx.\end{aligned}$$

To find the extremum we set the variation to 0:

$$\delta\widetilde{\mathcal{R}}[\mathbf{u}^*;\boldsymbol{\eta}] = 0.$$

By the fundamental lemma of the calculus of variations, this implies, for each $k$,

$$A(x)e^{u_k^*(x)/2} - B(x)p_k(x)e^{-u_k^*(x)/2} = 0.$$

Thus

$$A(x)e^{u_k^*(x)/2} = B(x)p_k(x)e^{-u_k^*(x)/2}.$$

Multiplying by $e^{u_k^*(x)/2}$ gives

$$A(x)e^{u_k^*(x)} = B(x)p_k(x),$$

and therefore

$$e^{u_k^*(x)} = p_k(x)\frac{B(x)}{A(x)}.$$

Taking logarithms,

$$u_k^*(x) = \log p_k(x) + \log \frac{B(x)}{A(x)}.$$

The second term is independent of the model index $k$. Hence the stationary solutions have the form

$$\boxed{u_k^*(x) = \log p(x \mid \mathcal{M}_k) + c(x)},$$

where $c(x)$ is a common, non-identifiable offset.

Taking the difference between any two outputs eliminates this common offset:

$$u_i^*(x) - u_j^*(x) = \log p(x \mid \mathcal{M}_i) - \log p(x \mid \mathcal{M}_j).$$

Therefore,

$$\boxed{u_i^*(x) - u_j^*(x) = \log \frac{p(x \mid \mathcal{M}_i)}{p(x \mid \mathcal{M}_j)} = \log K_{ij}(x)},$$

and removing the shorthand notation

$$\boxed{\mathcal{J}_\alpha(f_i^*(x)) - \mathcal{J}_\alpha(f_j^*(x)) = \log K_{ij}(x)}.$$

To verify that this stationary family gives the global minimum, note that at each $x$,

$$A(x)B(x) = \left(\sum_i p_i(x)e^{-u_i(x)/2}\right)\left(\sum_i e^{u_i(x)/2}\right).$$

By the Cauchy–Schwarz inequality,

$$A(x)B(x) \geq \left(\sum_i \sqrt{p_i(x)}\right)^2.$$

Equality holds if and only if

$$\sqrt{p_i(x)}\, e^{-u_i(x)/4} = \lambda(x)e^{u_i(x)/4}$$

for every $i$, which is equivalent to

$$u_i(x) = \log p_i(x) - 2\log \lambda(x) \equiv \log p_i(x) + c(x).$$

Thus the variational stationary family obtained above attains the pointwise lower bound of the population-risk integrand and is therefore a global minimizer. The common function $c(x)$ remains undetermined because the loss depends only on pairwise differences $u_i - u_j$.